Optimizing Splicing Junction Detection in Next Generation Sequencing Data on a Virtual-GRID Infrastructure

Olivier Terzo, Lorenzo Mossucca
Infrastructure and Systems for Advanced Computing (IS4AC)
Mario Boella Istitute (ISMB)
Via P.C. Boggio 61, Torino, Italy
Email: (terzo,mossucca)@ismb.it

Andrea Acquaviva, Francesco Abate,
Elisa Ficarra, Rosalba Provenzano
Department of Control and Computer Engineering
Politecnico di Torino
Torino, Italy
Email: (andrea.acquaviva, francesco.abate)@polito.it
elisa.ficarra@polito.it, rosalba.provenzano@studenti.polito.it

***Abstract*—The new protocol for sequencing the messenger RNA in a cell, named RNA-seq produce millions of short sequence fragments. Next Generation Sequencing technology allows more accurate analysis but increase needs in term of computational resources. This paper describes the optimization of a RNA-seq analysis pipeline devoted to splicing variants detection, aimed at reducing computation time and providing a multi-user/multi-sample environment. This work brings two main contributions. First, we optimized a well-known algorithm called TopHat by parallelizing some sequential mapping steps. Second, we designed and implemented a hybrid virtual GRID infrastructure allowing to efficiently execute multiple instances of TopHat running on different samples or on behalf of different users, thus optimizing the overall execution time and enabling a flexible multi-user environment.**

I. Introduction

Next Generation Sequencing (NGS) technologies represent a disruptive innovation in the field of functional genomic research leading to an unprecedented availability of biological data [1]. A NGS sequencer can produce millions of reads in a single run that must be then further processed to extract biological knowledge. The main novelty consists in the possibility of sequencing an entire genome or transcriptome with substantially lower costs and timings respect to the previous Sanger sequencing methodology [2]. The reason behind this biotechnological performance increase is the capability of NGS machines to chop the DNA/RNA molecules into small fragments, namely *reads*, that are successively sequenced in parallel with considerable saving in terms of time and economic resources. From a biological and technical point of view NGS technology leads to new challenges in terms of development of tools and computational infrastructures [5]. In fact, computing infrastructures and software tools must be able to handle the huge amount of sequencing data produced by biotechnological laboratories. This capability is key to perform, in a reasonable amount of time, those analysis related to the understanding of biological and pathological processes, such as gene expression profiling, small non coding RNA profiling, novel genes discovery, aberrant transcript event detection [3], [4].

For instance, reads alignment is a basic operation that maps the reads on a genome reference in order to reconstruct the original sample sequence. This step is used as building block of more complex analysis pipelines, such as for alternative splicing or gene fusion detection. However, due to the large number of reads, this basic block becomes critical. Several algorithms and the associated tools have been recently developed to optimize the alignment phase, in particular Bowtie [6] is widespread because of its effectiveness when dealing with short reads (50-100bp) with respect to previous solutions. Moreover, it supports multi-threaded processing and it makes an efficient usage of system memory. A more complex analysis flow is required to accomplish splicing detection, which is required to correctly reconstruct transcripts taking into account exon rearrangements (e.g. alternative splicing events), and thus to perform a more accurate alignment and expression profiling. Splicing detection requires to perform run the alignment algorithm in various phases of the analysis. Various algorithms and tools have been proposed to carry out this task. TopHat [7] is a splicing detection tool built on top of Bowtie, which is invoked within various execution loops. Moreover, TopHat presents many other processing steps performed in pipe to these alignments. Such a complex pipeline imposes tight requirements in terms of processing, memory utilization and storage space not only for input and output data but also for temporary files. Resource utilization is also dependent on the specific execution phase, thus imposing a dynamic behavior. To address these requirements and promote resource-efficient NGS data processing, parallel software implementations on scalable and flexible computing infrastructures are needed. In this work, we propose a new implementation of a splicing detection algorithm on a virtualized grid infrastructure which makes efficient use of CPU and memory also enabling a multi-user environment. In particular, we implemented a parallelized and modular version of TopHat, going beyond the inherent multithreading support of Bowtie, supporting multiple parallel sample processing and performing a careful evaluation of computation versus communication trade-off, which is critical when large data masses have to be transferred.

The infrastructure is an integrated system devoted to handle automatically DNA/RNA samples in multiuser context. It is composed of user-level tools (i.e. TopHat, Bowtie), middleware for the infrastructure management, central repository, relational database and specific scheduler for resource and job controller. The software used for the entire processing chain, installed on each worker node, is a modified version of native TopHat. In the new infrastructure, each parallelized TopHat instance runs to process a single sample. Many samples processing can be interleaved and the scheduler allocates the various execution steps on the physical resources, namely the worker nodes, to minimize execution time. The infrastructure leverages upon the integration of a hybrid and flexible computing infrastructure composed by physical workers nodes in a common standard grid computing architecture with virtual grid nodes in a cloud computing infrastructure.

The Globus Toolkit is used as middleware [8], since it allows obtaining a reliable information technology infrastructure that enables the integrated, collaborative use of computers, networks and databases. Two schedulers have been developed: Global and Local scheduler [9]. The Global Scheduler, installed on grid master, allows to collect all jobs to elaborate and to distributing them among the worker nodes. Instead, the Local Scheduler is installed on each worker nodes and its aim is to process data in accordance with the grid master request.

Experiments carried out to characterize the performance of the proposed solution show the effectiveness in reducing the execution time with respect to a standard version of TopHat and highlight the improved resource utilization allowed by the virtual infrastructure.

The rest of the paper is organized as follows: Sect. 2 presents the motivations, Sect. 3 gives an overview of the TopHat algorithm and Bowtie tools, Sect. 4 describes the computing infrastructure and how TopHat was adapted for a parallelization of some mapping reads process in a multi user environment, Sect. 5 shows performance aspects, Sect. 6 draws the conclusions and road map for future work.

II. Motivation

In NGS technology context the amount of data and the number of samples to be analyzed is growing constantly. It is a positive factor for increasing more accurate studies and results in term of reliable identifications of mutations in aberrant splicing events, fused genes and open new perspective and challenges for adapting tools that are in condition to make pre and post processing in modern infrastructure like virtualized machine, Grid and Cloud computing. A NGS data sample consists in a millions of data and the time needed for the process execution increase dramatically with only one workstation for processing NGS data. The alignment phase is a process which each mapping reference is made in independent way and can be execute in a parallel way on a distributed computing context. The alignment is a very basic operation but actually the alignment tool that we use, TopHat, provides a sequential analyze of each block of reads and in a single user way and consequently in a typical scenarios were more samples need to be analyze the total time for processing data is not acceptable. In consideration of this a first challenge is to adapt the tool by making a reverse engineering of the code for a transformation of all possible main and heavy sequential alignment process and to a parallelization way in order to obtain an optimization of process time. A second challenge is to adapt the alignment tool for an multi sample environment. This new scenario characterize the capabilities to make splice junction mapping in an flexible and distributed computing infrastructure.

III. Next Generation Sequencing

A. Tophat Algorithm

TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons. TopHat is a collaborative effort between the University of Maryland Center for Bioinformatics and Computational Biology and the University of California, Berkeley Departments of Mathematics and Molecular and Cell Biology. TopHat receives as input reads produced by the Illumina Genome Analyzer, although users have been successful in using TopHat with reads from other technologies [7]. The input samples consist of two file of about 37 million of reads each. The two file are FASTA formatted paired-end reads. Dealing with paired-end reads means that the reads are sequenced by the sequencing machine only on the end of the same DNA/RNA molecule, thus the sequence in the middle part is unknown. Each sequenced end of the same read is also referred as mate. It results in two distinct files, the first one consists in the first mate of the same reads and the second one consists in the opposite mate. TopHat finds junctions by mapping reads to the reference in two phases. In the first phase, the pipeline maps all reads to the reference genome using Bowtie. All reads that do not map to the genome are set aside as initially unmapped reads. Bowtie reports, for each read, one or more alignment containing no more than a few mismatches in the 5'-most s bases of the read. The remaining portion of the read on the 3' end may have additional mismatches, provided that the Phred-quality-weighted Hamming distance is less than a specified threshold.

B. Alignment Tools: Bowtie

The short reads alignment is surely the most common operation in RNA-Seq data analysis. The purpose of the alignment is to map each short read fragment onto a genome reference. From the computational point of view, each short read consists in a sequence of four possible characters corresponding to the DNA bases and the sequence length depends on the sequencing machine adopted for the biological experiment [6]. The main novelty introduced by NGS technology is the capability of sequencing small DNA/RNA fragments in parallel, increasing the throughput and producing very short reads as output. However, this feature make the computational problem more challenging because of the higher amount of read produced and

the accuracy in the mapping (the shorter sequence length, the higher probability of having multiple matches). For this reason many alignment tools specifically focussed on the alignment of short reads have been recently developed. In the present work, we are interested in characterizing the performances of alignment tools on real NGS data. On the wave of this remark, Bowtie has been chosen, a wide diffused alignment program particularly aimed at align short reads. In order to detect the actual limitation of the alignment phase, we consider real NGS data coming from the analysis of Chronic Myeloid Leukemia. In our analysis flow, the HG19 assembly produced in the 2007 is considered as reference genome the last human genome assembly produced to now. In order to increase the computational performances during the read mapping, Bowtie program creates an index of the provided human genome reference. This operation is particularly straightforward from the computational point of view, but it must be performed only one time for the human genome reference and it is independent on the mapping samples. The alignment phase itself is particularly suitable to be parallelized. In fact, each mapping operation is applied to each read independently on the other read mapping.

IV. VirtualBIO NGS Infrastructure

A. Overview

In a preliminary phase of reverse engineering, studying TopHat, blocks of transactions have been highlighted that were executed sequentially. We have identified three main blocks(see Figure 1), that can be executed independently:

a left and right check reads segments;
b left and right mapping segments with HG19;
c left and right mapping segments with segment juncs.

A feature of these 3 blocks is that they are performed by a external software (Bowtie). For steps (a) and (c), since the files involved in the development are significant, we created a common repository that contains the temporary folder used by TopHat. Instead the step (b) uses small files these can be performed on a grid, both physical and virtual, because the transfer times are lower. Only difference that the input files are transferred to worker node through Globus. These worker nodes when the process is terminated, re-send the output file to the node that requested execution. This platform aims to be a service that is given to biologists for the NGS analysis but not only, the solution also provides a case study where multiple users require the execution of analysis simultaneously. The architecture, called VirtualBIO NGS (see Figure 2), is composed of three main components: a Master Node(MN), a part consists of the Physical Worker Nodes (PWN) that set the grid environment while a part consists of Virtual Worker Nodes (VWN) that set the virtualized environment [8]. The MN is a physical machine with good hardware characteristics, is responsible of Certification Authority, contains the database, where all information about the nodes belonging to the infrastructure, have stored, the node status, the flow of the various biological analysis that can be made in the system

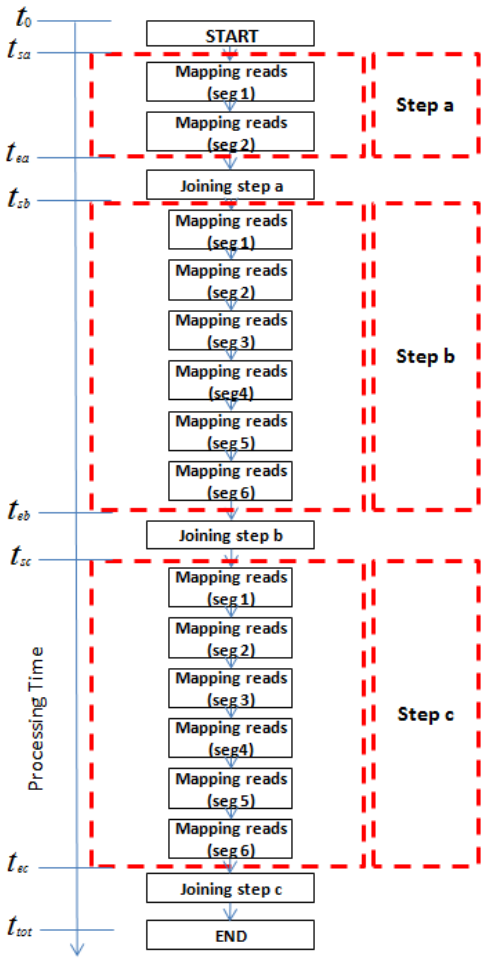

Fig. 1. Simple Tophat Flow.

and system monitoring. Both environments are configured with the middleware Globus Toolkit, since it allows obtaining a reliable information technology infrastructure that enables the integrated, collaborative use of computers, networks and databases. The Globus Toolkit is a collection of software components designed to support the development of applications for high performance distributed computing environments, or computational grids.

B. VirtualBIO NGS Distributed Environment

The grid environment consists of machines with high computing power. Grid environments are scalable, making them effective for uses where storing large amounts of data are important. The only requirement is to have the necessary software installed for the processing (Bowtie and Globus). In each worker node of the grid environment is installed the Grid Local Scheduler, an essential component for performing biological tests. Virtualized environment also help to improve infrastructure management, allowing the use of virtual node template to create virtual nodes in a short time, speeding up the integration of new nodes on the grid and, therefore, improving the reactivity and the scalability of the infrastructure.

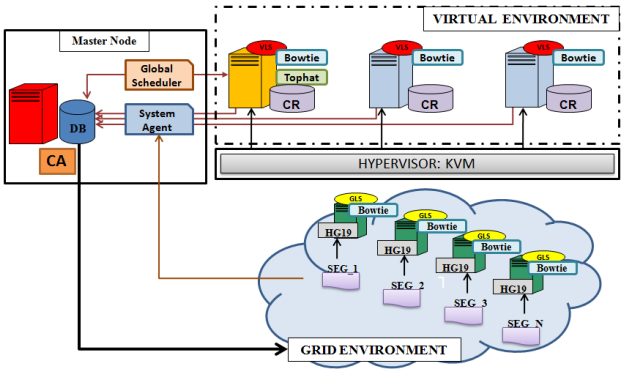

Fig. 2. Computing Architecture.

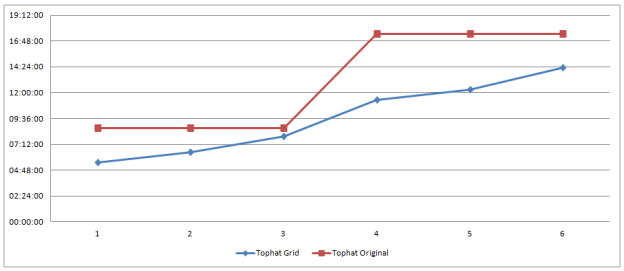

Fig. 3. Original Tophat vs Tophat Grid.

Hypervisor KVM was used for the creation of Full Virtualized machines. By adding virtualization capabilities to a standard Linux kernel, the virtualized environment can benefit from all the ongoing work on the Linux kernel itself. Under this model, every virtual machine is a regular Linux process, scheduled by the standard Linux scheduler. The virtualized environment has pre-installed images, which contain all software(Bowtie and TopHat), local schedulers (VLS, GLS) and support data (HG19). To have images already configured allows to set up easily machines when you need them, and once used the close the instance.

1) Scheduler Context: The Local Grid Scheduler (GLS) is a scheduler active on physical machines, has been developed for the design phase (b), it aligns the segment with respect to the human genome (HG19) through Bowtie. Since the transfer of the input file is not influential, the worker nodes do not need to be in the same subnet as the master node, but may also belong to different virtual organization, so system can have greater scalability and can use machines powerful performance [9]. The Virtual Local Scheduler (VLS) is a scheduler active on virtual machines. Its purpose is to draw up the steps (a) and (c) of TopHat. As the GLS, the VLS performs the mapping files for input received through Bowtie. The step (a) allows the alignment with respect to the human genome (HG19) and step (c) allows the alignment with respect to the segments junc previously constituted by TopHat. Since the considerable size of the files involved in these 2 steps, the VLS works directly on the temporary folder that is located in the common repository, allowing to avoid wasting time due to the transfer of data. Even in this case the interaction with the database is essential and very frequent, network problems may affect the entire biological analysis.

V. Performances

The aim of tests performed is to compare execution time for multi samples using the two version of TopHat: the original version that sequential approach and version modified exploiting the distributed environment. For this test phase, we wanted to use an architecture which consists of six machines with four CPUs. In Figure 3, we can notice that already only a sample processed with the version of TopHat grid, a time savings of 40% is obtained instead increasing the number of samples to be processed, it is worth noting that the percentage of earned time is about 30%, this is due to the jobs queues that are created on the nodes.

VI. Conclusions and Future Works

VirtualBIO NGS is a tool for NGS analysis, in particular for the alignment phase through TopHat and Bowtie. The solution covered both the field of infrastructure and the optimization software. Infrastructure is based on grid and virtual environment, using a common repository and a couple of job scheduler. The TopHat algorithm has been optimized making parallel independent sections that were sequential and has been modified for giving a multi user environment were before on the native version of TopHat was for a single user instance. The system allows to reduces the elaboration time for a single sample by at least 30% allowing to use own machines but not only, this value can change it depends on the power of the machines used. Future works include the improvement of scheduling policies, balancing jobs and resources, this study also opens to a scenario for increasing the capabilities of the scalability of the infrastructure through an integration of Cloud machines hosted in Amazon.

References

[1] De Magalhes JP, Finch CE, Janssens G., *Next-generation sequencing in aging research: Emerging applications, problems, pitfalls and possible solutions.*, Ageing Research Reviews, 2010 Jul;9(3):315-23.

[2] F. Sanger, S. Nicklen, A.R. Coulson, *DNA sequencing with chain-terminating inhibitors*, Proc. Natl. Acad. Sci. USA 74 (1977) 54635467.

[3] Kircher M, Kelso J., *High-throughput DNA sequencing concepts and limitations.*, Bioessays. 2010 Jun;32(6):524-36. Review.

[4] Maher CA, Palanisamy N, Brenner JC, Cao X, Kalyana-Sundaram S, Luo S, Khrebtukova I, Barrette TR, Grasso C, Yu J, Lonigro RJ, Schroth G, Kumar-Sinha C, Chinnaiyan, *Chimeric transcript discovery by paired-end transcriptome sequencing*, AM. Proc Natl Acad Sci U S A. 2009 Jul 28

[5] M. Pop, S.L. Salzberg, *Bioinformatics challenges of new sequencing technology*, Trends Genet. 24, 2008

[6] Langmead B, Trapnell C, Pop M, Salzberg SL. *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome*. Genome Biology 10:R25.

[7] Trapnell C, Pachter L, Salzberg SL. *TopHat: discovering splice junctions with RNA-Seq*. Bioinformatics doi:10.1093/bioinformatics/btp120

[8] Berman, Fox, Hey, *Grid Computing Making the Global Infrastructure a Reality*, Wiley,2005

[9] Kurowsky, Nabrzyzki, *Scheduling jobs on the Grid-Multicriteria approach*, Computational Methods in Science and Technology, 2006